建立食品卫生安全保障体系数学模型及改进模型的若干理论问题

1. 问题简述

从我国国情出发，通过建立人群食物摄入量模型、污染物分布模型、风险评估模型，构建我国食品卫生安全保障体系。

其中，人群食物摄入量模型（膳食模型）是用于估计不同地区、不同性别、不同年龄、不同季节、不同劳动强度、不同经济收入的人群各类食品的一天摄入量；污染物分布模型是根据农药、化工等污染行业的污染物排放数据和食品卫生安全监测部门日常对水、农贸市场和大宗食品中污染物的抽查数据以及进出口口岸的检测数据来估计各类食物中各种污染物的含量；风险评估模型则根据前两个模型所提供的数据计算得出全国或某地区人群某些污染物每天摄入量的99.999%的右分位点，从而能够对某一时刻食品安全风险作出评估。

要求建立基本模型并改进模型中的若干理论问题，具体工作如下：

人群食物摄入量模型中，要考虑：

1. 如何设计抽样调查方案使调查结果能尽量反映全国的实际情况，调查结果的数据使用起来效果比较理想，同时使调查的全部工作量在可以承受的范围内；
2. 中国居民消费的食品种类比其他国家居民消费的食品种类复杂得多，包括：主食、肉类、蔬菜、水果、水、饮料、各种调味剂和经过加工的食品，细分将达数千种以上，在实际调查过程中进行如此详细地分类，其调查工作量太大，而如果随意粗糙进行分类，则将影响调查的精度，因此需要根据污染物分布模型的数据合理设计抽样调查中食物的分类办法；
3. 要用通过万分之一（甚至更小）的抽样率得到的数据建立起全国比较准确的人群食品摄入量模型，因此要确定合理的技术路线，充分利用从其他一切渠道可以获得的信息，可以并且应该建立不止一个这样的模型以满足各方面的需求。

污染物分布模型中，要考虑：

1. 我们应该怎样充分利用这批抽样率很低的数据去建立模型；
2. 由于食品的季节性、区域性、多样性特点，日常监测无法获得详细的、完整的分类数据，问题是如何利用这些数据尽量提高模型的精度；
3. 由于监测时间方面的要求和经费的限制，在日常检测时往往采用比较快捷的检测方法，即符合性检验，其缺点是当检测项目的检测结果是安全时就不再精确测量污染物具体的含量了，而笼统地用“未检出”作为检测结果。这对判断这批食品是否安全而言是完全满足要求的，但作为污

染物分布模型的输入而言，如果“未检出”全部当成零来计算就一定会产生比较大的误差，因此一定要改进。

风险评估模型，即利用前两个模型的结果对全国、某个地区、某类食品的安全状况做出评价，对可能出现的食品安全事件给出预警。其中，要考虑：

1. 如何处理数据不配套问题，即人群食品摄入量模型中的调查对象极大可能不是污染物分布模型中被调查食品的消费者；
2. 前两个模型的数据分类也很可能不配套，人群食品摄入量模型中的食品很可能远多于污染物分布模型中被调查食品或者两者的分类不完全一致；
3. 模型要求给出全体居民某项污染物摄入量的 99.999%的右分位点，那么模型采用什么方法提高它的精度。

除上所述，还要考虑：尽量提出有创造性的技术路线并解决大量理论问题，如：污染物含量的分布呈左偏态、抽样样本不配套、统计分类标准不同、由信息贫乏的样本估计全国的总体情况等问题。

2. 问题假设

1) 某地区、全国人群食物摄入量服从正态分布；
2) 食品中污染物含量服从对数正态分布；
3) 日常对市场上食物的检测看作在食品出厂前的检测；
4) 运输途中污染物的增量和运输时间成正比；

3. 基本符号说明

p_j——表示第 j 个地区抽取的城镇数；

q_j——表示第 j 个地区抽取的乡村数；

Ac_{ijkl}——第 i 类食品在第 j 个地区的第 k 个城市随机抽取的第 l 户家庭的人均一天摄入量；

Av_{ijkl}——第 i 类食品在第 j 个地区的第 k 个农村随机抽取的第 l 户家庭的人均一天摄入量；

B_{ik}——市场上第 i 类食物中第 k 种污染物的含量；

Bs_{ik}——国内未经过运输的第 i 类食物第 k 种污染物的含量；

Ds_{ik}——进口的第 i 类食物中第 k 种污染物的含量；

C_{ik}——第 i 类食物中第 k 种污染物在流通中随机扰动变化量；

G_{ijk}——第 j 个地区人群在一天内摄入第 i 类食物中第 k 类污染物含量；

4. 模型建立与求解

4.1 人群食物摄入量模型

4.1.1 问题分析

需要解决的主要问题：

1. 某个地区和全国人群对第i类食品的一天摄入量的概率分布，
2. 设计合理的抽样方案使其能尽量反应全国的实际情况
3. 建立全国的比较准确的人群食品摄入量模型，作为理论问题在后面进行研究

4.1.2 抽样调查设计

由于统计中存在的抽样方法很多，因此我们在选择抽样方法的时候需要考虑如何提高精度以及节省成本。由于我国人口众多，地域特色明显，从节省人力物力以及经费方面来考虑，显然直接在全国范围内进行简单随机抽样是不可行的，而整群抽样在通常情况下抽样误差比较大，所以也不予采用。而根据地域特色等因素，我们在本文中，是把国家划为不同区域，随机抽取一些样本地区，由于在实际生活中，城乡的饮食习惯存在着较大的差异，因此这里我们把这些样本地区分为城乡两层，紧接着在这些城乡中分别独立随机抽取一些样本进行入户调查。因此我们采用的是分层抽样与多阶段抽样的结合，而多阶段抽样是保持了整群抽样的样本比较集中，便于调查，节省经费等优点，同时还避免了对小单元进行过多调查所造成的浪费，而分层抽样的估计精度比较高，因此我们选择这两种抽样方法的结合。这样给出不同食物不同地区不同城乡的家庭人均摄入量就可以较为精确的估计某种食物在不同地区的人均摄入量，甚至可以得出某种食物在全国的家庭人均摄入量。在假设这个摄入量服从正态分布的时候，我们可以得出摄入量的分布。

首先采用分层抽样，将我国化为R个不同区域，并从中随机抽取r个地区。由于城乡饮食存在较明显的差异，故分别针对这r个地区，将每个地区的人群划分为：城镇层和乡村层，在这里是把全国的某类食物人均摄入量看成是一个总体，则这个总体有R个单元，再将这R个单元分成两个层，每个层的大小是分别是Rc,Rv，每一层的样本量分别是rc,rv。接着采用二阶段抽样，然后分别在这些已经随机抽出来的地区中随机抽取城镇与乡村，假设在第j个地区随机抽取p_j个城镇与q_j个乡村进行调查，而在不妨假设每个地区的城镇有P_j个，乡村Q_j个，因为这些数据是可以在相关部门查到的。最后再在抽出来的城镇和乡村中，随机抽取一些家庭进行入户调查，不妨假设在城镇中选取n_1个家庭入户调查，在乡村中选取n_2个家庭入户调查。

4.1.3 人群食物摄入量模型的建立与求解

我们假设某地区及全国的人群某种食物A摄入量均服从正态分布，可以运用矩估计方法来确定未知参数——均值μ及方差σ^2。

4.1.3.1 各地区人群食物摄入量的估计

通过抽样，得到样本数据：

Ac_{ijk}——第i类食品在第j个地区的第k个城市的人群人均一天摄入量集合

Av_{ijk}——第i类食品在第j个地区的第k个农村的人群人均一天摄入量集合

根据这些统计量可以估计出某地区人群食物人均摄入量所服从的分布的参数，进而得到它相应服从的分布，具体过程如下：

首先，我们可以根据两组样本数据，以及多阶段的估计方法，得到如下估计量。

第i类食物在第j个地区的第k城镇家庭的人均摄入量的样本均值：

$$\overline{A}c_{ijk}=\frac{1}{n_1}\sum_{k=1}^{n_1}Ac_{ijkl}\qquad(i=1,2,...,s;j=1,2,...,r;k=1,2,...,p;l=1,2,...,n_1)$$

其中s表示食品的种类总数，r表示调查地区总数，k某地区的城镇总数，n_1表示样本的个数。

第i类食物在第j个地区的城镇家庭人均摄入量的样本均值：

$$\overline{A}c_{ij}=\frac{1}{p_j}\frac{1}{n_1}\sum_{k=1}^{p_j}\sum_{l=1}^{n_1}Ac_{ijkl}$$

第i类食物在第j个地区的第k城镇家庭的人均摄入量的样本方差：

$$s_1(Ac_{ijk})=\frac{1}{p_j}\frac{1}{n_1-1}\sum_{l=1}^{n_1}\left(Ac_{ijkl}-\overline{A}c_{ijk}\right)^2$$

第i类食物在第j个地区城镇家庭的人均摄入量的样本方差：

$$s_2(Ac_{ij})=\frac{1}{p_j-1}\sum_{k=1}^{p_j}(\overline{A}c_{ijk}-\overline{A}c_{ij})^2$$

由此不难得出：

$$v(Ac_{ij})=\frac{1-f_1}{p}s_1^2+\frac{f_1(1-fc_2)}{pn_1}s_2^2,$$

其中，$f_1=\frac{p_j}{P_j},fc_2=\frac{n_1}{N_1}$

这样得出第i类食物在第j个地区的城镇家庭人均摄入量服从的分布为：$N(\overline{A}c_{ij},v(Ac_{ij}))$。

同理可得，农村的各个相应估计量：

第i类食物在第j个地区的第k乡村的人均摄入量的样本均值：

$$\overline{A}v_{ijk}=\frac{1}{n_2}\sum_{k=1}^{n_2}Av_{ijkl}\qquad(i=1,2,...,s;j=1,2,...,r;k=1,2,...,q;l=1,2,...,n_2)$$

第i类食物在第j个地区的乡村家庭人均摄入量的样本均值：

$$\overline{A}v_{ij}=\frac{1}{q_j}\frac{1}{n_2}\sum_{k=1}^{q_j}\sum_{l=1}^{n_2}Av_{ijkl}$$

第i类食物在第j个地区的第k个乡村家庭的人均摄入量的样本方差：

$$s_1(Av_{ijk})=\frac{1}{q_j}\frac{1}{n_2-1}\sum_{l=1}^{n_2}\left(Av_{ijkl}-\overline{A}v_{ijk}\right)^2$$

第i类食物在第j个地区乡村家庭的人均摄入量的样本方差：

$$s_2(Av_{ij})=\frac{1}{q_j-1}\sum_{k=1}^{q_j}(\overline{A}v_{ijk}-\overline{A}v_{ij})^2$$

由此不难得出：

$$v(Av_{ij})=\frac{1-f_1}{q_j}s_1^2+\frac{f_1(1-fv_2)}{q_jn_2}s_2^2,$$

其中，$f_1=\frac{q_j}{Q_j},fv_2=\frac{n_2}{N_2}$

这样得出了第i类食物在第j个地区的农村人均摄入量服从的分布为：$N(\overline{A}v_{ij},v(Av_{ij}))$。

由上述这些估计量，将某个地区的某类食物的摄入量看成是一个总体，则这个总体有$Rc+Rv$个单元，再将这$Rc+Rv$个单元分成两个层，每个层的大小是分别是Rc,Rv，因此，我们可以根据分层抽样的方法来计算第i类食物在第j个地区的相应估计量。

故求出第i类食物在第j个地区的人均一天摄入量的样本均值与方差：

$$\overline{A}_{ij}=\frac{Rc}{Rv+Rc}\overline{A}c_{ij}+\frac{Rv}{Rv+Rc}\overline{A}v_{ij}$$

$$v(A_{ij})=\frac{Rc}{Rc+Rv}\frac{1-fc}{p_j}s_2(\overline{A}c_{ij})+\frac{Rv}{Rc+Rv}\frac{1-fv}{p_j}s_2(Av_{ij})$$

从而，第i类食物在第j个地区的人群一天摄入量服从正态分布，即：

$$A_{ij}\sim N\left(\overline{A}_{ij},v(A_{ij})\right)$$

4.1.3.2 全国人群食物摄入量估计

在得出了第i类食物在第j个地区的人均一天摄入量的样本均值与方差后，我们再来进一步估计全国人群的第i种食物人均摄入量。

第i种食物在全国范围内人均摄入量的样本均值与方差分别如下：

$$\overline{A}_i=\frac{1}{r}\sum_{j=1}^{r}\overline{A}_{ij}$$

$$s^2=\frac{1}{r-1}\sum_{j=1}^{r}(\overline{A}_{ij}-\overline{A}_i)^2$$

因为我们前面假设全国人群某类食物摄入量服从正态分布，根据上面的式子整理后就可以得到正态分布中参数的估计量，即我们可以得到正态分布的未知参数的近似估计值：

$$\hat{\mu}_i = \overline{A}_i$$

$$\hat{\sigma}_i^2 = v(A_i)$$

这样即得全国人群某类食物摄入量的分布，即 $A_i \sim N(\overline{A}_i, v(A_i))$。

4.2 污染物分布模型

4.2.1 问题分析

需要解决的主要问题：

1. 根据污染物的含量来确定食物的分类问题
2. 低随机抽样数据的处理问题，如何提高精度
3. 截尾数据的处理

4.2.2 食物的分类问题

我国居民消费的食品种类纷繁杂乱，包括：主食、肉类、蔬菜、水果、水、饮料、各种调味剂和经过加工的食品，细分将达数千种以上，在实际调查过程中进行如此详细地分类，其调查工作量太大，而如果随意粗糙进行分类，则将影响调查的精度，因此，我们首先需要根据污染物含量的数据，合理地将食物分类。

通过相关资料，我们不难发现：含有某类污染物的食物种类是有限的，例如：有机磷多存在于蔬菜水果中，二噁英主要存在于动物源性食品中，如：蛋类、肉、水产和奶等。

因此，我们在检测某种污染物含量后，可根据该污染物含量的不同对各类食物做聚类分析，将具有相近污染物含量的不同食物聚为一类，这样就可以有效缩减我国食物繁杂种类的数目。而且，经过食物分类的统一还可以避免样本分类不配套的问题。

（注：下面的讨论，是在建立在食物由此方法已分类的基础上进行的。）

4.2.3 污染物分布的简单分析

由于污染物的含量分布不是正态分布，而是偏向于左偏态，因此我们假设它是服从对数正态分布的，然后因为市场上出售的食物，大致是由两部分组成的，一部分是本国出产，一部分是进口的。而食物要能在市场上出售就必然经过运输途径，因此不可避免在运输途中食物有可能受到其他的污染物污染，也可能是食物温度升高或降低而导致的食物内部元素含量产生变化。因此，我们将市场上食物的污染物含量的检测样本看成是两个独立的样本构成，即：食物出厂或刚收获时污染物的检验和进出口污染物的检验，而把运输途中污染物的变化看成是一个对它们产生影响的随机扰动项。

根据上述原因，我们在进行抽样时，分别对国内没经过运输的食物（即刚出厂的食物或者刚收获的食物等）以及进口的食物的污染物含量进行随机抽样，然后把运输途中污染物含量的变化看成是服从正态分布的扰动。在这里，为了简便处理，把运输途中的污染物改变量单独提出来考虑，因为我们在对进口食物进行

抽样时，也是在关口进行检查的（即在国内还尚未流通）。

4.2.4 关键符号的再说明

B_{ik}——市场上第i类食物中第k种污染物的含量，服从对数正态分布

Bs_{ik}——国内未经过运输的第i类食物第k种污染物的含量，服从对数正态分布

Ds_{ik}——进口的第i类食物中第k种污染物的含量，服从对数正态分布

C_{ik}——第i类食物中第k种污染物在流通中随机扰动变化量，服从正态分布

4.2.5 污染物分布基本模型的建立与求解

根据我们前面关于污染物含量分布的假设，由抽得的样本数据，同样用矩估计来得出各个分布的参数，如下：

$$\hat{\mu}_{Bs_{ik}} = \frac{1}{m_1}\sum_{\ell=1}^{m_1} \ln Bs_{ik\ell}$$

$$\hat{\sigma}^2_{Bs_{ik}} = v(\ln Bs_{ik}) = \frac{1}{m_1 - 1}\sum_{\ell=1}^{m_1} (\ln Bs_{ik\ell} - \hat{\mu}'_{Bs_{ik\ell}})^2$$

$$\hat{\mu}_{Ds_{ik}} = \frac{1}{m_2}\sum_{\ell=1}^{m_2} \ln Ds_{ik\ell}$$

$$\hat{\sigma}^2_{Ds_{ik}} = v(\ln Ds_{ik}) = \frac{1}{m_2 - 1}\sum_{\ell=1}^{m_2} (\ln Ds_{ik\ell} - \hat{\mu}'_{Ds_{ik}})^2$$

可以得到：$\ln Bs_{ik} \sim N(\hat{\mu}_{Bs_{ik}}, \hat{\sigma}^2_{Bs_{ik}})$，$\ln Ds_{ik} \sim N\ (\hat{\mu}_{Ds_{ik}}, \hat{\sigma}^2_{Ds_{ik}})$，$C_{ik} \sim N(\hat{\mu}_{cik}, \hat{\sigma}^2_{cik})$。

因为它们都是服从对数正态分布的，所以其对数是服从正态分布的。这样，就与扰动项同属于一种分布。为了简便处理，我们不妨假设运输过程中的污染物含量的变化与运输时间成正比，比例系数λ，即它的正态分布的未知参数可看作关于运输时间的函数，由此，可以得出：

$$\hat{\mu}_{ik} = \lambda \cdot t$$

我国当地出产的食物与进口的食物在进入市场出售时，必然要经过运输，所以在上述两者基础上加上运输所导致的扰动项，即是在两者原来的基础上进行修正，这样就可以得到市场上出售的食物污染物含量。

在考虑扰动项时，首先不管是本国出产的还是在市场上出产的都应该经过运输，因此在抵达市场时，它们原有的污染物含量都会产生不同程度的变化。在这里不妨假设这个扰动项只是起到一个修正作用，因此应该在它们原来的污染物含量的基础上再加上运输途中污染物含量的变化，即得到它们流通到市场上后污染物的含量。由于扰动项只是起到一个修正作用，所以这样并不会改变它们的总体分布。根据独立正态分布具有可加性，而且，本国的食物污染物含量与进口食物的污染物含量是相互独立，是显而易见的。所以，修正后的参数可以用下面的式子表示：

$$\hat{\mu}'_{Bs_{ik}} = \frac{1}{m_1}\sum_{\ell=1}^{m_1}\ln(Bs_{ik\ell} + \hat{\mu}_{C_{ik}})$$

$$\hat{\mu}'_{Ds_{ik}} = \frac{1}{m_2}\sum_{\ell=1}^{m_2}\ln(Ds_{ik\ell} + \hat{\mu}_{C_{ik}})$$

$$\hat{\sigma}'^2_{Bs_{ik}} = v(\ln Bs_{ik} + \hat{\mu}_{C_{ik}}) = \frac{1}{m_1 - 1}\sum_{\ell=1}^{m_1}(\ln(Bs_{ik\ell} + \hat{\mu}_{C_{ik}}) - \hat{\mu}'_{Bs_{ik}})^2$$

$$\hat{\sigma}'^2_{Ds_{ik}} = v(\ln Ds_4 + \hat{\mu}_{C_{ik}}) = \frac{1}{m_2 - 1}\sum_{\ell=1}^{m_2}(\ln(Ds_{ik\ell} + \hat{\mu}_{C_{ik}}) - \hat{\mu}'_{Ds_{ik}})^2$$

这里，由于对本国食物的选择与进口食物抽样的样本量不相等，而二者又是独立服从对数正态分布的，而市场的食物又主要是由这两部分构成的，所以市场食物也是服从对数正态分布，并且它的参数可以由下面的式子获得：

市场上第 i 类食物的第 k 类污染物的含量的样本均值估计值：

$$\hat{\mu}'(\ln B_{ik}) = \alpha\hat{\mu}'_{Bs_{ik}} + (1-\alpha)\hat{\mu}'_{Ds_{ik}}$$

市场上第 i 类食物的第 k 类污染物的含量的样本方差估计值：

$$\hat{\sigma}'^2(\ln B_{ik}) = \alpha^2\hat{\sigma}'^2_{Bs_{ik}} + (1-\alpha)^2\hat{\sigma}'^2_{Ds_{ik}}$$

这里，由于两种抽样的样本量不同，因此，我们采用加权来估算市场上食物的污染物含量，在这里我们可以把本国的食物污染物含量与进口污染物的含量看成是两种市场上食物污染物含量的两个不同的层，因此根据分层抽样，我们可以得到，α 实际就是层权，在这里假设我们抽样的时候大致与现实生活中的层权相同，即我们所随机抽取的的两个样本量的比例与现实生活中总体的比例相差不大，因此，我们可以近似的用每个样本量与所抽取的总的样本量之比来估计层权，层权就可以表示成如下形式：

$$\alpha = w_1 = \frac{m_1}{m_1 + m_2},$$

$$1 - \alpha = w_2 = \frac{m_2}{m_1 + m_2}$$

由此可得：

$$\hat{\mu}'(\ln B_{ik}) = \alpha\hat{\mu}'_{Bs_{ik}} + (1-\alpha)\hat{\mu}'_{Ds_{ik}} = \frac{m_1}{m_1 + m_2}\hat{\mu}'_{Bs_{ik}} + \frac{m_2}{m_1 + m_2}\hat{\mu}'_{Ds_{ik}}$$

$$\hat{\sigma}'^2(\ln B_{ik}) = \alpha^2\hat{\sigma}'^2_{Bs_{ik}} + (1-\alpha)^2\hat{\sigma}'^2_{Ds_{ik}} = (\frac{m_1}{m_1 + m_2})^2\hat{\sigma}'^2_{Bs_{ik}} + (\frac{m_2}{m_1 + m_2})^2\hat{\sigma}'^2_{Ds_{ik}}$$

此时，已经得到了市场上第 i 类食物中第 k 种污染物的含量 $\ln B_{ik}$ 服从的正态分布是 $N(\hat{\mu}'(\ln B_{ik}), \hat{\sigma}'^2(\ln B_{ik}))$，所以当我们得到抽样数据后很容易根据上面的方法计算出市场上第 i 类食物中第 k 种污染物含量的分布 $B_{ik} \sim LN(\overline{\mu}_{ik}, \overline{\sigma}^2_{ik})$，其中 $\overline{\mu}_{ij} = \hat{\mu}'(\ln B_{ik})$，$\overline{\sigma}^2_{ik} = \hat{\sigma}'^2(\ln B_{ik})$。

4.2.4 模型的改进（考虑截断数据）

根据题目可知：由于检测时间方面的要求及经费的限制，在日常的检测中往往采用比较简便快捷的检测方法，即符合性检验，当检测项目的检测结果是安全时就不再精确测量污染物的具体含量了，而是笼统地用“未检出”作为检测结果。

统计学中的可靠性理论对此已有相关研究，将上述数据情况称为截断数据，但其只对右截尾情况有比较成熟的研究，而本问题中是针对左截尾的情况。

4.2.4.1 截断数据相关概念

从一个服从分布函数的总体中，随机抽取 n 个样本，如果在这 n 个样本中大于某个特定值的样本观测之是缺省时，那么这组样本的观测值称为右截尾数据。反之，称为左截尾数据。 [3]

4.2.4.2 “未检出”数据的插补

作为污染物分布模型的输入而言，如果“未检出”全部当成零来计算就一定会产生比较大的误差，尤其是当食品大多数属于合格，即大多数样本数据以“未检出”作为结果时，若不对数据做相应处理，则在对总体分布的估计中，将产生无法估量的误差。因此我们对数据做如下改进：

我们可以利用占抽查数据的 2%偶然抽样调查数据，来补充小于等于截断值（作为未检出的临界含量值）的部分抽样数据。从偶然抽样调查数据中，可以获得 N_2 个小于等于截断值的抽样数据。利用这 N_2 抽样数据随机模拟出 N_1 个数据，来作为定时截断的数据值($N_2 < N_1$)。这里我们给出两种成熟的模拟方法，即：

1. 蒙特卡罗模拟
2. BP 网络仿真

具体算法[7]在这里就不再赘述，但模拟时应注意到：当模拟时出现大于截断值的数据时，我们将其剔除。

4.2.4.3 改进模型的建立与求解

不妨设随机变量 X 服从对数正态分布，其概率密度函数为：

$$p_X(x)=\frac{1}{\sigma x\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right),\quad x>0$$

在抽样过程中，小于某一给定值 τ 的抽样数据为失效数据，为定时左截尾问题，设样本量为 N，利用 4.2.4.2 给出的方法将 N_1 个失效数据进行补全，其值为 x_i，$i=1,\cdots,N_1$，我们需要给出未知参数 μ，σ 估计值 $\left(\hat{\mu},\hat{\sigma}^2\right)$。

目前对对数正态分布 $LN\left(\mu,\sigma^2\right)$ 定时左截尾情形下，还没有估计未知参数 μ，σ 的有效的方法，我们考虑将其转化成一个随机变量的右截尾问题来求解，而对一个随机变量的右截尾问题，可以通过回归方法或者极大似然方法进行参数的估计。

考虑随机变量$Y=\frac{1}{X}$，通过函数变换$y=\frac{1}{x}$，我们可以将随机变量X的左截尾参数估计问题转化为随机变量Y的右截尾参数估计问题。当抽样数据值$x_i<\tau$时，抽样数据为左截尾数据，此时$y_i=\frac{1}{x_i}>\frac{1}{\tau}$为右截尾数据，在抽样过程中，随机变量$Y$的截断值为$\frac{1}{\tau}$，样本容量为$N$，失效数据的个数为$N_1$，其值为$y_i=\frac{1}{x_i}$

首先，我们要确定随机变量$Y=\frac{1}{X}$的概率分布：$y=\frac{1}{x}(x>0)$是严格单调的，且其反函数$x=h(y)=\frac{1}{y}(y>0)$存在连续的导函数，从而$Y=\frac{1}{X}$的概率密度函数为：

$$p_Y(y)=p_X(h(y))|h(y)|=\frac{y}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln 1/y-\mu)^2}{2\sigma^2}\right)\frac{1}{y^2} \qquad y>0$$
$$=\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(-\frac{(\ln y+\mu)^2}{2\sigma^2}\right)$$

分布函数为：

$$F_Y(y)=\int_{-\infty}^{y}p_Y(y)dy=\int_0^y\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(-\frac{(\ln y+\mu)^2}{2\sigma^2}\right)dy$$

对连续随机变量右截尾问题的极大似然函数，我们有如下的结论：

定理 1：X是一任意的一维连续随机变量，其概率密度函数为$f(x;\theta_1,\cdots,\theta_k)$，分布函数为$F(x;\theta_1,\cdots,\theta_k)$，则在考虑右截尾数据分析时的极大似然函数为：

$$L(\theta_1,\cdots,\theta_k\mid x_1,\cdots,x_n,\tau)=\frac{N!}{(N-n)!}\prod_{i=1}^{n}f(x_i,\theta_1,\cdots,\theta_n)[1-F(\tau,\theta_1,\cdots,\theta_n)]^{N-n}$$

其中：N为样本数据，n为失效数据的个数，k为被估计的未知参数的个数，

$\theta_i(i=1,\cdots,k)$为被估计的未知参数，$x_i(i=1,\cdots,k)$为失效数据的值

利用此命题,随机变量Y右截尾问题的极大似然函数为：

$$L(\mu,\sigma\mid y_1,\cdots,y_{N_1},\tau)=\frac{N!}{(N-N_1)!}\prod_{i=1}^{N_1}p_Y(y_i;\mu,\sigma)[1-F_Y(\tau;\mu,\sigma)]^{N-N_1}$$
$$=\frac{N!}{(N-N_1)!}\prod_{i=1}^{N_1}\frac{1}{\sigma y_i\sqrt{2\pi}}\exp\left(-\frac{(\ln y_i+\mu)^2}{2\sigma^2}\right)[1-F_Y(\tau;\mu,\sigma)]^{N-N_1}$$

其中，$F_Y(\tau;\mu,\sigma)=\int_0^{\tau}\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(-\frac{(\ln y+\mu)^2}{2\sigma^2}\right)dy$

则：

$$\ln L\left(\mu,\sigma \mid y_1,\cdots,y_{N_1},\tau\right)=\ln\frac{N!}{\left(N-N_1\right)!}+\sum_{i=1}^{N_1}\ln\frac{1}{\sigma y_i\sqrt{2\pi}}\exp\left(-\frac{\left(\ln y_i+\mu\right)^2}{2\sigma^2}\right)+\left(N-N_1\right)\ln\left[1-F_Y\left(\tau;\mu,\sigma\right)\right]$$

$$=\ln\frac{N!}{\left(N-N_1\right)!}-\sum_{i=1}^{N_1}\ln\sigma-\sum_{i=1}^{N_1}\ln y_i\sqrt{2\pi}-\sum_{i=1}^{N_1}\frac{\left(\ln y_i+\mu\right)^2}{2\sigma^2}+\left(N-N_1\right)\ln\left[1-F_Y\left(\tau;\mu,\sigma\right)\right]$$

从而：

$$\begin{cases}\dfrac{\partial\ln L\left(\mu,\sigma \mid y_1,\cdots,y_{N_1},\tau\right)}{\partial\mu}=-\displaystyle\sum_{i=1}^{N_1}\frac{\ln y_i+\mu}{\sigma^2}-\frac{N-N_1}{1-F_Y\left(\tau;\mu,\sigma\right)}\cdot\frac{\partial F_Y\left(\tau;\mu,\sigma\right)}{\partial\mu}\\ \dfrac{\partial\ln L\left(\mu,\sigma \mid y_1,\cdots,y_{N_1},\tau\right)}{\partial\sigma}=\displaystyle\sum_{i=1}^{N_1}\frac{\left(\ln y_i+\mu\right)^2}{4\sigma^3}-\frac{N_1}{\sigma}-\frac{N-N_1}{1-F_Y\left(\tau;\mu,\sigma\right)}\cdot\frac{\partial F_Y\left(\tau;\mu,\sigma\right)}{\partial\sigma}\end{cases}$$

其中，

$$\frac{\partial F_Y\left(\tau;\mu,\sigma\right)}{\partial\mu}=\frac{\partial}{\partial\mu}\int_0^\tau\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(-\frac{\left(\ln y+\mu\right)^2}{2\sigma^2}\right)dy=-\int_0^\tau\frac{\ln y+\mu}{\sigma^2}\cdot\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(-\frac{\left(\ln y+\mu\right)^2}{2\sigma^2}\right)dy$$

$$\frac{\partial F_Y\left(\tau;\mu,\sigma\right)}{\partial\sigma}=\frac{\partial}{\partial\sigma}\int_0^\tau\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(-\frac{\left(\ln y+\mu\right)^2}{2\sigma^2}\right)dy=-\int_0^\tau\frac{4\sigma^2-\left(\ln y+\mu\right)^2}{4\sigma^3}\cdot\frac{1}{\sigma y\sqrt{2\pi}}\exp\left(-\frac{\left(\ln y+\mu\right)^2}{2\sigma^2}\right)dy$$

令方程组满足：

$$\begin{cases}\dfrac{\partial\ln L\left(\mu,\sigma \mid y_1,\cdots,y_{N_1},\tau\right)}{\partial\mu}&=0\\ \dfrac{\partial\ln L\left(\mu,\sigma \mid y_1,\cdots,y_{N_1},\tau\right)}{\partial\sigma}&=0\end{cases}$$

则：该方程组的解，即为我们所要估计的参数$\left(\hat{\mu},\hat{\sigma}\right)$。由于此方程组为复杂的非线性方程组，无法求出其解析解，所以，只能通过数值方法进行求解。

4.3 风险评估模型：

4.3.1 问题分析

需解决的理论问题：

1. 数据不匹配：所调查食物摄入量的对象与所调查食物的消费者不匹配
2. 数据分类不匹配：调查食物摄入量的食物种类与被检测的食物种类不匹配
3. 提高右分位点的计算精度

4.3.2 模型建立与求解

在食品的风险评估中，暴露模型是一个很重要的评估方法。所用到的食品消费与市场上各种食品中污染物的存在及其含量这两大类数据之间没有直接的关系。在暴露评估中也没有一个数据能够代表所有个体的摄入量以及摄入食物中污染物的浓度。因此，食品的安全的暴露评估经常需要建立模型来代表真实的暴露情况。广义而言，代表这种暴露的模型可以用下面的公式来表示：消费量×某污染物的含量=膳食暴露。

由于在风险评估模型中，某类污染物的暴露量=摄入量×食品中某类污染物

的含量[5]（*）。

当人群食品摄入量模型中的调查对象与污染物分布模型中被调查食品的消费者为同一人群时，可以利用（*）式进行风险评估，但实际抽样调查中，人群食品摄入量模型中的调查对象极大可能不是污染物分布模型中被调查食品的消费者，从而存在抽样数据不配套的问题，不能直接利用（*）式进行风险评估。

下面我们给出几种处理的方法：

1. 简化人群食物摄入量模型：

假设：所有个体对某一特定食品的摄入水平是一样的。

在这一假设的基础上，我们可以考虑利用点评估模型计算暴露量，进而做风险评估分析。

点评估模型或决定论模型包括在一个模型中对每一个评估参数使用一个单一的“最佳猜想”，这一方法中，需进一步假定：从各种来源的暴露量等于食品摄入量（如平均的或较高的摄入量数据）的固定值乘以污染物含量（通常是平均含量或国家允许值的上限）。

利用这一方法，我们利用调查对象各类食品的一天摄入量的平均值来近似代替消费人群各类食品一天摄入量的平均值，也可以将食品中污染物的含量用于调查人群，从而可以对调查人群及消费人群进行风险评估。

进一步，我们利用调查人群各类食品的一天摄入量，估计出了某一地区人群各类食品的一天摄入量，将污染物的分含量应用于这一人群，就可以进行风险评估分析。

具体说，根据人群食物摄入量模型，我们此时已知：

A_{ij} ——第 i 类食品在第 j 个地区人群一天摄入量，服从正态分布 $N\left(\mu_{ij},\sigma_{ij}^2\right)$

B_{ik} ——第 i 类食品中第 k 种污染物的含量，服从对数正态分布 $LN\left(\overline{\mu}_{ik},\overline{\sigma}_{ik}^2\right)$

则第 j 个地区第 k 类污染物的暴露量 G_{jk} 可以表示为：

$$G_{jk}=\sum_{i=1}^{s}\mu_{ij}\overline{\mu}_{ik}$$

其中一天摄入量的均值 μ_{ij} 也可以用一个较高的摄入量数据代替，污染物含量的均值 $\overline{\mu}_{jk}$ 也可以用国家允许值的上限来代替

由此我们可以看出点估计存在以下的缺点：

1）这一方法不能对在一个人群中可能发生的，或影响评估结果的主要因素等所有可能的暴露提供一个预警。

2）当高水平的值用于代替人群的摄入量均值或污染物浓度均值时，把多种食物的摄入量综合起来就会导致评估过高甚至常常不切实际，为精确评估暴露，需要有能够整合食品摄入量与污染物浓度的更为复杂的方法，这样才能更切实际的反应出真实的暴露情况。

3）无法给出某污染物摄入量的右分位数。

针对上述缺点，我们可以看出点估计精度不够高，所以下面我们给出点评估的一个改进方法：

我们仍假设：所有个体对某一特定食品的摄入水平是一样，但食物中某类污染物的含 B_{ik} 用分布 $N\left(\overline{\mu}_{ik},\overline{\sigma}_{ik}^2\right)$ 来表示，此时

$$G_{jk}=\sum_{i=1}^{s}\mu_{ij}B_{ik}$$

这是一个随机变量，根据实际情况，可认为各 B_{ik} 之间是相互独立的，利用对数正态分布的可加性，此时 G_{jk} 服从分布

$$G_{jk}\sim LN\left(\sum_{i=1}^{s}\mu_{ij}\overline{\mu}_{ik},\sum_{i=1}^{s}\mu_{ij}^2\overline{\sigma}_{ik}^2\right)$$

改进以后，考虑了污染物含量的变化情况，表达的是污染物分布的情况，比点估计能够获得更具有价值的信息，此外还可以计算某污染物摄入量的右分位数。

2. **简化污染物分布模型：**

假定：某特定食品中某类污染物的浓度为一固定值。

在这一假设的基础上，我们可以考虑利用单一分布模型进行风险评估。

暴露评估中的单一分布模型表达的是人群各类食品一天摄入量的分布，对污染物的含量使用一个固定的参数值来表达。这样，当某一地区的人群各类食品一天摄入量的分布知道以后，就可以对此人群进行风险评估。此时第 k 类污染物的暴露量可以表示为：

$$G_{jk}=\sum_{i=1}^{s}\overline{\mu}_{ik}A_{ij}$$

这也是一个随机变量，根据实际情况，可认为各 A_{ij} 之间是相互独立的，利用正态分布的可加性，此时 G_{jk} 服从分布：

$$G_{jk}\sim N\left(\sum_{i=1}^{s}\overline{\mu}_{ik}\mu_{ij},\sum_{i=1}^{s}\overline{\mu}_{ik}^2\sigma_{ij}^2\right)$$

显然，此时也可以给出某污染物摄入量的右分位数。

3. 概率分析模型

上面两种处理方法中，都做了一定的假设，这对问题的分析带来了方便，但同时引起的误差比较大，下面我们考虑从概率分布的角度利用抽样调查数据进行风险评估。

概率分析包括各种参数变化性与不确定性对总体分布的影响，它通过发生的概率对模型中的每一个参数可能引发的各种结果进行考虑。概率分析用于食品中污染物的的风险评估分析，用来描述食品污染物的暴露风险分布。在食品污染物

的暴露概率分析的模型中，食品摄入数据及食品中污染物的分布均采用概率分布，并且依据每一个输入的分布，找出与暴露过程相一致的数学模型，再随机生成一些数据来进行污染物的暴露。

前面我们已经获得：

A_{ij}——第i类食品在第j个地区人群的一天摄入量，服从正态分布$N\left(\mu_{ij},\sigma_{ij}^2\right)$

B_{ik}——第i类食品中第k种污染物的含量，服从对数正态分布$LN\left(\overline{\mu}_{ik},\overline{\sigma}_{ik}^2\right)$

如果知道第j个地区第i类食品的一天摄入量的分布A_{ij}，则

$$G_{jk}=\sum_{i=1}^{s}A_{ij}B_{ik}$$

如果不知道第j个地区第i类食品的一天摄入量A_{ij}的分布，则全国人群第i类食品的一天摄入量A_i的分布$N\left(\mu_i,\sigma_i^2\right)$，并用$A_i$来代替第$j$个地区人群第$i$类食物的摄入量，此时：

$$G_{jk}=\sum_{i=1}^{s}A_iB_{ik}$$

关于$G_{jk}=\sum_{i=1}^{s}A_{ij}B_{ik}$及$G_{jk}=\sum_{i=1}^{s}A_iB_{ik}$的概率分布的求法将在后面给出。

下面以第j个地区为例，来推导$G_{jk}=\sum_{i=1}^{s}A_{ij}B_{ik}$的概率分布（$G_{jk}=\sum_{i=1}^{s}A_iB_{ik}$的情况可类似推导）

首先推导$A_{ij}B_{ik}$的概率分布：

根据实际情况，我们可以假设随机变量A_i与B_{ij}是相互独立的，则人群第i类食品一天的摄入量A_{ij}的概率密度函数为：

$$p\left(A_{ij},x\right)=\frac{1}{\sigma_{ij}\sqrt{2\pi}}\exp\left(-\frac{\left(x-\mu_{ij}\right)^2}{2\sigma_{ij}^2}\right)\quad,（x>0）$$

第i类食品中第k类污染物的含量B_{ik}的概率密度函数为：

$$p\left(B_{ik},x\right)=\frac{1}{\sigma_{ik}y\sqrt{2\pi}}\exp\left(-\frac{\left(\ln x-\mu_{ik}\right)^2}{2\sigma_{ik}^2}\right),（x>0）$$

则因此人群在一天内摄入第i类食物中第k类污染物含量$G_{ijk}=A_{ij}B_{ik}$的概率密度函数可以表示为：

$$p(G_{ijk},x)=\int_{-\infty}^{\infty}p(A_{ij},x/v)p(B_{ik},v)\frac{1}{|v|}dv$$

$$=\int_{-\infty}^{\infty}\frac{1}{\sigma_{ij}\sqrt{2\pi}}\exp\left(-\frac{(x/v-\mu_{ij})^2}{2\sigma_{ij}^2}\right)\cdot\frac{1}{\overline{\sigma}_{ik}v\sqrt{2\pi}}\exp\left(-\frac{(\ln v-\overline{\mu}_{ik})^2}{2\overline{\sigma}_{ik}^2}\right)\cdot\frac{1}{|v|}dv \qquad (x>0)$$

$$=\frac{1}{2\pi\sigma_{ij}\overline{\sigma}_{ik}}\int_0^{\infty}\frac{1}{v}\exp\left(-\frac{(x/v-\mu_{ij})^2}{2\sigma_{ij}^2}-\frac{(\ln v-\overline{\mu}_{ik})^2}{2\overline{\sigma}_{ik}^2}\right)dv$$

根据上式可以给出，右分位数 x_p 的确定方法如下：

$$\frac{1}{2\pi\sigma_{ij}\overline{\sigma}_{ik}}\int_0^{x_p}\frac{1}{v}\exp\left(-\frac{(x/v-\mu_{ij})^2}{2\sigma_{ij}^2}-\frac{(\ln v-\overline{\mu}_{ik})^2}{2\overline{\sigma}_{ik}^2}\right)dv=1-p$$

这里，我们给出计算 x_p 的数值方法的算法如下，流程图见图 4-1。

Step1. 给出 x_p 的一个初始估计值 x_o，通过数值积分的方法计算积分值：

$$F(x_0)=\frac{1}{2\pi\sigma_{ij}\overline{\sigma}_{ik}}\int_0^{x_o}\frac{1}{v}\exp\left(-\frac{(x/v-\mu_{ij})^2}{2\sigma_{ij}^2}-\frac{(\ln v-\overline{\mu}_{ik})^2}{2\overline{\sigma}_{ik}^2}\right)dv$$

Step2. 比较 $F(x_0)$ 与 $1-p$，若 $F(x_0)<1-p$，转 Step3；若 $F(x_0)>1-p$，转 Step4。

Step3. while$(F(x_0)<1-p)$ do $\{x_0+\Delta x\to x_0$ ； 计算积分值$F(x_0)\}$

$x_0-\Delta x$ 为所求的分位点, 程序结束。

Step4. while$(F(x_0)>1-p)$ do $\{x_0-\Delta x\to x_0$ ； 计算积分值$F(x_0)\}$

x_0 为所求的分位点, 程序结束。

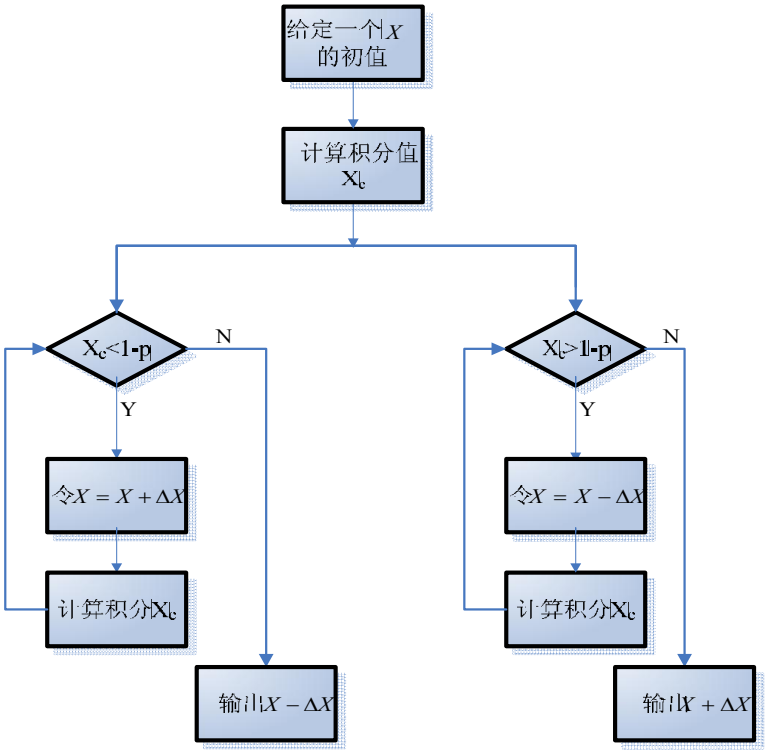

图 4-1

根据实际情况，可以认为 $A_{ij}B_{ik}$ ， $i=1,\cdots,s$ 是相互独立的，由此可以利用连续场合的卷积公式计算 G_j 的概率分布：

为此，先考虑只有两类食物含同种污染物的情况，即： $G_{jk}=G_{1jk}+G_{2jk}=A_{1j}B_{1k}+A_{2j}B_{2k}$ 的概率分布情况：

已知： $p\left(G_{1jk},x\right)=\dfrac{1}{2\pi\sigma_{1j}\overline{\sigma}_{1k}}\int_0^\infty\dfrac{1}{v_1}\exp\left(-\dfrac{\left(x/v_1-\mu_{1j}\right)^2}{2\sigma_{1j}^2}-\dfrac{\left(\ln v_1-\overline{\mu}_{1k}\right)^2}{2\sigma_{1j}^2}\right)dv_1$

$$p\left(G_{2jk},x\right)=\frac{1}{2\pi\sigma_{2j}\overline{\sigma}_{2k}}\int_0^\infty\frac{1}{v_2}\exp\left(-\frac{\left(x/v_2-\mu_{2j}\right)^2}{2\sigma_{2j}^2}-\frac{\left(\ln v_2-\overline{\mu}_{2k}\right)^2}{2\overline{\sigma}_{2k}^2}\right)dv_2$$

则： $p\left(G_{1jk}+G_{2jk},x\right)=\int_{-\infty}^{\infty}p\left(G_{1jk},x/w_1\right)p\left(G_{2jk},w_1\right)dw_1$

$$\int_{-\infty}^{\infty}\frac{1}{\left(2\pi\right)^2\sigma_{1j}\sigma_{2j}\overline{\sigma}_{1k}\overline{\sigma}_{2k}}\int_0^\infty\int_0^\infty\frac{1}{v_1v_2}\exp\left(\begin{array}{l}-\dfrac{\left(x/w_1v_1-\mu_{1j}\right)^2}{2\sigma_{1j}^2}-\dfrac{\left(w_1/v_2-\mu_{2j}\right)^2}{2\sigma_{2j}^2}\\-\dfrac{\left(\ln v_1-\overline{\mu}_{1k}\right)^2}{2\overline{\sigma}_{1k}^2}-\dfrac{\left(\ln v_2-\overline{\mu}_{2k}\right)^2}{2\overline{\sigma}_{2k}^2}\end{array}\right)dv_1dv_2dw_1$$

=

$$==\frac{1}{\left(2\pi\right)^2\sigma_{1j}\sigma_{2j}\overline{\sigma}_{1k}\overline{\sigma}_{2k}}\int_0^\infty\int_0^\infty\int_0^\infty\frac{1}{v_1v_2}\exp\left(\begin{array}{l}-\dfrac{\left(x/w_1v_1-\mu_{1j}\right)^2}{2\sigma_{1j}^2}-\dfrac{\left(w_1/v_2-\mu_{2j}\right)^2}{2\sigma_{2j}^2}\\-\dfrac{\left(\ln v_1-\overline{\mu}_{1k}\right)^2}{2\overline{\sigma}_{1k}^2}-\dfrac{\left(\ln v_2-\overline{\mu}_{2k}\right)^2}{2\overline{\sigma}_{2k}^2}\end{array}\right)dv_1dv_2dw_1$$

接着我们考虑有三类食物含同种污染物的情况，即： $G_{jk}=G_{1jk}+G_{2jk}+G_{3jk}=\left(G_{1jk}+G_{2jk}\right)+G_{3jk}$ 的概率分布情况：

增加已知条件： $p\left(G_{3jk},x\right)=\dfrac{1}{2\pi\sigma_{3j}\overline{\sigma}_{3k}}\int_0^\infty\dfrac{1}{v_3}\exp\left(-\dfrac{\left(x/v_3-\mu_{3j}\right)^2}{2\sigma_{3j}^2}-\dfrac{\left(\ln v_3-\overline{\mu}_{3k}\right)^2}{2\overline{\sigma}_{3k}^2}\right)dv_3$

则：

$$p\left(G_{1jk}+G_{2jk}+G_{3jk},x\right)=\int_{-\infty}^{\infty}p\left(G_{1jk}+G_{2jk},x/w_2\right)p\left(G_{3jk},w_2\right)dw_2$$

$$=\int_{-\infty}^{\infty}\left[\begin{array}{l}\dfrac{1}{(2\pi)^2\sigma_{1j}\sigma_{2j}\bar{\sigma}_{1k}\bar{\sigma}_{2k}}\int_0^{\infty}\int_0^{\infty}\int_0^{\infty}\dfrac{1}{v_1v_2}\exp\left(\begin{array}{l}-\dfrac{\left(x/w_1w_2v_1-\mu_{1j}\right)^2}{2\sigma_{1j}^2}-\dfrac{\left(w_1/v_2-\mu_{2j}\right)^2}{2\sigma_{2j}^2}\\-\dfrac{\left(\ln v_1-\bar{\mu}_{1k}\right)^2}{2\bar{\sigma}_{1k}^2}-\dfrac{\left(\ln v_2-\bar{\mu}_{2k}\right)^2}{2\bar{\sigma}_{2k}^2}\end{array}\right)dv_1dv_2dw_1\\\cdot\dfrac{1}{2\pi\sigma_{3j}\bar{\sigma}_{3k}}\int_0^{\infty}\dfrac{1}{v_3}\exp\left(-\dfrac{\left(w_2/v_3-\mu_{3j}\right)^2}{2\sigma_{3j}^2}-\dfrac{\left(\ln v_3-\bar{\mu}_{3k}\right)^2}{2\bar{\sigma}_{3k}^2}\right)dv_3\end{array}\right]dw_2$$

$$=\frac{1}{(2\pi)^3\sigma_{1j}\sigma_{2j}\sigma_{3j}\bar{\sigma}_{1k}\bar{\sigma}_{2k}\bar{\sigma}_{3k}}\int_0^{\infty}\int_0^{\infty}\int_0^{\infty}\int_0^{\infty}\int_0^{\infty}\frac{1}{v_1v_2v_3}$$

$$\exp\left(\begin{array}{l}-\dfrac{\left(x/w_1w_2v_1-\mu_{1j}\right)^2}{2\sigma_{1j}^2}-\dfrac{\left(w_1/v_2-\mu_{2j}\right)^2}{2\sigma_{2j}^2}-\dfrac{\left(w_2/v_3-\mu_{3j}\right)^2}{2\sigma_{3j}^2}\\-\dfrac{\left(\ln v_1-\bar{\mu}_{1k}\right)^2}{2\bar{\sigma}_{1k}^2}-\dfrac{\left(\ln v_2-\bar{\mu}_{2k}\right)^2}{2\bar{\sigma}_{2k}^2}-\dfrac{\left(\ln v_3-\bar{\mu}_{3k}\right)^2}{2\bar{\sigma}_{3k}^2}\end{array}\right)dv_1dv_2dv_3dw_1dw_2$$

由此依次类推可得 $H_j=\sum_{i=}^{s}H_{ij}=\sum_{i=1}^{s}A_iB_{ij}=A_1B_{1j}+A_2B_{2j}+\cdots+A_sB_{sj}$ 的概率分布。

在所求得的这样一种复杂的概率分布时，右 x_p 分位数的确定方法为：采用数值计算的方法，虽然理论上是可行的，但积分计算时困难很大。

5 改进模型的若干理论问题研究

5.1 低抽样率情况下总体概率分布的估计问题

也就是说要利用利用若干个低抽样率的样本的概率分布比较精确地估计估计总体的概率分布，这个问题可以描述为：已知总体 X 的 n 个样本 $X_1,\cdots,X_n$ 及相应的概率分布 $F_1(x_1),\cdots,F_n(x_n)$，如何比较精确地估计 X 的概率分布？

思路一：

假设总体 X 的概率分布 $F(x)$ 为各个样本的概率分布 $F_1(x),\cdots,F_n(x)$ 的线性加权组合，加权系数为样本的抽样比，此种假设下，我们认为各个样本之间是相互独立的，这与实际情况往往是不符合的，而且由于抽样率很低，我们有必要考虑各个样本之间的关联性，以便获得更多的信息，来计较精确地估计总体的概率分布。

思路二：

假设总体 X 的概率分布可看成是样本 $X_1,\cdots,X_n$ 的联合分布，$F_1(x),\cdots,F_n(x)$ 为此联合概率分布的 n 个边缘分布，也就是把总体 X 的概率分布看作是 n 维联合分布，问题转化成已知边缘分布，如何确定联合分布？在此种假设下，我们考虑了各个样本之间的关联性，联合分布函数中包含有更多的信息。

Copula 理论中的 Sklar 定理指出：对于一个具有一元边际分布的联合分布函

数$F(x_1,\cdots,x_n)$，一定存在一个 Copula 函数C，使得

$$F(x_1,\cdots,x_n)=C(F_1(x_1),\cdots,F_n(x_n))$$

如果$F_1,\cdots,F_n$是连续的，则C是唯一的；否则C不唯一。反之，如果C是n维 Copula 函数，$F_1,\cdots,F_n$为边缘分布函数，则如上定义的$F(x_1,\cdots,x_n)$是n分布函数。我们可以通过构造合适 Copula 函数，来确定联合概率分布。

人群食品摄入量模型中提出：如何用通过万分之一（甚至更小）的抽样率得到的数据建立起全国比较精确的人群食品摄入量模型？

利用抽样数据，我们已经给出了各个地区的人群食品摄入量的概率分布，具体说，第j个地区第i类食品人群一天摄入量A_{ij}满足分布$N(\mu_{ij},\sigma_{ij}^2)$，$j=1,\cdots s$，则全国人群第$i$类食品一天摄入量可以用这$s$个地区的联合分布来表示，利用 Sklar 定理，全国人群第i类食品一天摄入量服从下面的分布：

$$F(X)=F(x_1,\cdots,x_n)=C(F_1(x_1),\cdots,F_n(x_n))$$

其中，$F_i(x_i)$为正态分布$N(\mu_{ij},\sigma_{ij}^2)$的分布函数。

当边缘分布均为正态分布时，我们有如下的定理：

定理 2：假设$(X_1,\cdots,X_n)$服从多元正态分布，当且仅当其边缘分布函数$F_1,\cdots,F_n$皆为正态分布，存在唯一的 Copula 函数（Normal Copula），使得

$$C^N(\mu_1,\cdots,\mu_n)=\Phi(\phi^{-1}(\mu_1),\cdots,\phi^{-1}(\mu_n))$$

其中，Φ为多元标准正态分布的分布函数，ϕ^{-1}为一元标准正态分布函数的反函数。$\mu_1,\cdots,\mu_n$为 Copula 函数的参数[6]。

利用上面的定理，只要确定 Copula 函数$C^N(\mu_1,\cdots,\mu_n)=\Phi(\phi^{-1}(\mu_1),\cdots,\phi^{-1}(\mu_n))$中的参数就可以确定分布函数$F(X)=F(x_1,\cdots,x_n)=C(F_1(x_1),\cdots,F_n(x_n))$

此外，Copula 理论不限制边缘分布的选择，也就是说当边缘分布函数的类型不相同时，我们也可以通过构造 Copula 函数来确定联合分布函数，关于 Copula 函数的具体构造方法有待于进一步的深入研究。

5.2 随机变量存在截尾时分布函数或均值的估计问题

也就是说我们只知道随机变量在大于或小于某一数值的部分统计数据，如何才能有效估计此随机变量的分布函数或者均值？

若随机变量分布类型未知，还找不到一种用部分大于或小于某一数值的统计数据来估计其分布函数的方法，问题有待于进一步研究；若已知随机变量服从的分布，则问题转化为某一分布存在定时左截尾或右截尾时参数的估计问题。

污染物分布模型中，我们假设污染物的分布满足对数正态分布，前面我们已经给出了对数正态分布存在定时左截尾数据时，均值及方差的估计方法。对于一般分布存在左截尾数据时的处理问题，我们提出一种解决的思路：

不妨设随机变量X的概率密度函数为$f_X(x;\theta_1,\cdots,\theta_k)$

1) 构造变换函数 $y=g(x)$，将随机变量 X 转化为随机变量 Y，保证 X 左截尾数据为 Y 的右截尾数据，同时求出 Y 的概率密度函数 $f_Y(y;\theta_1,\cdots,\theta_k)$，进而可以得到 Y 的分布函数 $F_Y(y;\theta_1,\cdots,\theta_k)$

2) Y 的极大似然估计函数为

$$L(\theta_1,\cdots,\theta_k \mid y_1,\cdots,y_n,\tau)=\frac{N!}{(N-n)!}\prod_{i=1}^{n} f(y_i,\theta_1,\cdots,\theta_n)[1-F(\tau,\theta_1,\cdots,\theta_n)]^{N-n}$$

3) 利用极大似然估计法对未知参数 $\theta_1,\cdots,\theta_k$ 进行估计

5.3 关于保障食品安全的一些想法

食品卫生安全保障体系的中的一项重要工作是建立模型以刻画食品的质量，我们考虑建立一套科学、合理的食品质量安全综合评价指标体系并采用先进的技术对其进行评估。通过观察食品安全综合评价指标值可以发现不同时期食品安全存在的问题，食品安全综合评价指标中的各项基本指数又是随着自然条件、经济发展及农业发展等多种不确定因素的影响而发生明显的变化，研究在现实中通过科学、高效的技术方法判断哪些主要指标严重影响着食品安全的某个方面，哪些指标对食品安全影响程度较大，具有很重要的意义。

在实际的抽样调查中，考虑到检测时间方面的要求和经费等对抽样样本量的限制，往往抽样率都是很低的，如何利用这些数据进行建模并尽量提高模型的精度，是一个相当重要的问题，而灰色系统理论是从小样本、贫信息不确定系统中寻求规律，着重研究外延明确、内涵不明确的对象，采用从系统内部出发去发掘信息并充分利用信息建模。

在建立食品安全综合评价指标体系时，指标体系系统一般具有明显的层次结构，层次构成可以设为：

目标层：代表某一国家、地区或家庭的食品安全总体水平状况，是衡量食品安全水平高低的综合指标，可用0～1之间的数据表示，数值越接近于1，说明食品安全的综合水平越高，反之越低，其取值由下一些指标计算确定

指标层：是综合评价指标体系中最基础性的评价指标，可以从本质上反应食品安全在某环节中的状况，具有可测，可比及可获得性。

从而我们考虑采用层次分析模型和灰色关联模型来评判食品的安全性，具体的做法是：

1、确定各层指标的权重：构造判断矩阵，进行层次排序及一致性检验

2、各层指标的灰色关联分析：包括指标值的确定及无量纲处理；参考方案指标序列的分析；指标层指标值的标准化处理；指标层指标值的关联度计算

3、备选方案的综合关联度分析

该方法的优缺点分析：

对处理抽样率低的随机抽样数据能够获得比较高的精度，且能够比较客观、合理地评价食品的安全现状，又可大大提高工作效率和质量；但此种方法无法求出食品中各类污染物的概率分布情况，只是给出了食品的一个整体安全水平。

6. 模型评价

本文在建立人群食品摄入量模型时，采用由于分层与多阶段抽样想结合的方式进行抽样，提高了相对精度并节省了大量的经费。并因为采用这样一种抽样方式，使得充分利用样本信息，相对准确地估计各个地区和全国的人群食品摄入量成为了可能。

针对污染物分布模型，为了解决食物在运输到市场上的过程中污染物的含量会产生变化，我们把运输途中污染物含量看成是服从正态分布的扰动项，因此我们考虑市场上食物污染物含量的时候（分为本刚出产的尚未经过运输的食物与进口食物两部分组成），而在这里将这两部分别加上扰动项，作为混合样本，这样即可得到市场上污染物含量分布。由于考虑了运输过程中污染物的含量变化对市场上食物的污染物含量的影响，那么对总体分布的估计起到一个修正作用，使估计更精确。而当数据存在左截尾情况时，我们是先采用蒙特卡罗模拟或者 BP 网络仿真来补全出缺失数据，再根据可靠性理论中的定时截断来处理，就可以比较精确的得到市场上食物污染物含量的分布。这些都是本文的主要工作和亮点。

在风险评估模型中，主要解决的是模型的输入数据不匹配问题。我们分别采用了点评估模型、单一分布模型、概率分布模型来处理这个问题，并给出了某类污染物摄入量的概率分布。在此基础上给出了求解 99.999%的右分位点方法，遗憾的是没有给出提高右分位点精度的方法。

最后在改进模型时进行了若干理论问题研究，诸如低抽样率情况下总体概率分布的估计问题和随机变量存在截尾时分布函数或均值的估计问题，给出了解决了这个两个问题的试探性想法，但是否可行还有待商榷。

参考文献

【1】金勇进编著，抽样技术[M].中国人民大学出版社，2002，168～182

【2】茆诗松编著，统计手册[M].科学出版社，2003，767～774

【3】茆诗松等编著，可靠性统计[M].华东师范大学出版社，1984，267～284

【4】陈家鼎编著，生存分析与可靠性引论[M].1993，44～51

【5】罗祎等，论食品安全暴露评估模拟模型[J].中国检验检疫科学研究院，2007-10-22

【6】肖璨著，基于 copula 方法的二元组合风险模型与应用研究[硕士学位论文]，2007

【7】张恒喜等编著，小样本多元数据分析方法及应用[M].西北工业大学出版社，2002，154~161